Monitoreo y evaluación de transacciones financieras de alto impacto en volúmenes reducidos

# Parte 1

# Comprensión del Negocio

Este proyecto tiene como una premisa clave el monitoreo en tiempo real de transacciones en el sector financiero, especialmente ante eventos de indisponibilidad del servicio, dado su impacto directo en la experiencia del cliente.

Dentro del ecosistema transaccional, existe un conjunto de transacciones que, aunque no representan un volumen diario significativo, tienen una alta relevancia debido a su impacto potencial en el cliente y en las pérdidas financieras. En este contexto, resulta pertinente implementar un modelo analítico basado en machine Learning que permita un monitoreo eficiente, incluso en escenarios con volúmenes de datos reducidos

# Comprensión de los Datos

Los datos empleados en este estudio provienen de un histórico transaccional consolidado en dos bases generales del grupo Bancolombia, diseñadas para centralizar la información operativa y estratégica de los canales de atención y transacción del banco. Estas fuentes son parte de un Repositorio Transaccional de Canales, un entorno robusto de almacenamiento que ofrece trazabilidad, calidad de datos y cobertura integral de los eventos transaccionales.

El conjunto de datos tiene más de 20 millones de registros correspondientes al primer cuatrimestre del año 2025.

Este repositorio se compone de dos esquemas clave:

**Tabla en tiempo real (formato Kudu):**

Es la fuente principal de datos utilizada para la analítica en este proyecto. Al estar diseñada en formato Kudu, un sistema de almacenamiento columnar optimizado para grandes volúmenes de datos con baja latencia, esta tabla permite realizar consultas rápidas sobre información transaccional reciente. Su capacidad de integración con motores analíticos distribuidos como Apache Impala permite habilitar análisis en tiempo real, lo cual es crucial para la detección inmediata de anomalías en operaciones de baja volumetría.

**Tabla en procesamiento batch (formato Parquet):**

Complementariamente, se cuenta con una tabla histórica en formato Parquet, un formato columnar altamente eficiente para el procesamiento de grandes volúmenes de datos en modo batch. Esta fuente permite realizar análisis retrospectivos, identificar patrones de comportamiento y establecer líneas base históricas que serán utilizadas para la comparación con las señales actuales.

Parquet es un formato abierto de almacenamiento columnar optimizado para grandes volúmenes de datos y ampliamente usado en ecosistemas Big Data por su eficiencia en compresión y lectura

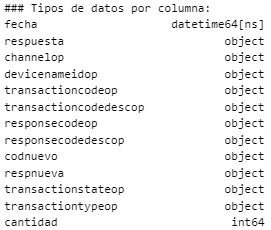
**Objetivo del uso combinado de ambas tablas**

El uso de estas dos estructuras en conjunto proporciona una ventaja analítica significativa. Mientras la tabla en Kudu habilita la evaluación en tiempo real del comportamiento de las transacciones actuales, la tabla en Parquet permite enriquecer el modelo con un contexto histórico robusto, facilitando así la identificación de desviaciones relevantes y la personalización de los umbrales de alerta.

Ejemplo de la Información sobre la metadata de las tablas que se usaran para el análisis de los datos: (Anexo – Tabla completa)

|  |  |  |
| --- | --- | --- |
| **Campo** | **Tipología** | **Descripción** |
| Fecha | FECHA | Fecha en el que se ingesta el registro de la transacción dentro de la base de datos. (Fecha de la realización de la transacción) |
| Canal | STRING | Nombre del canal transaccional por el cuál se ha realizado la transacción. |
| Código de Transacción | INT | Identificador numérico que representa el tipo específico de transacción realizada. Permite clasificar y filtrar operaciones según su naturaleza. |
| Respuesta de la transacción | STRING | Texto que indica el resultado de la transacción (por ejemplo, 'Exitosa', 'Fallida', Técnicamente exitosa). Es clave para las casuísticas del fallo de la transacción. |
| Cantidad | INT | Cantidad de transacciones, relacionada a la variable de Fecha. |

Algunos de los campos más importantes para el estudio son: 'fecha', 'respuesta', ‘Canal’, ' Código de Transacción ', Respuesta de la transacción’.



Tipos de datos de la tabla final.

**Análisis exploratorio.**

Como parte del estudio exploratorio para entender el comportamiento de las transacciones de bajo volumen, se realizó un análisis de series de tiempo que permitió descomponer la información en sus principales componentes: tendencia, estacionalidad y autocorrelación. Esta aproximación fue clave para identificar patrones repetitivos y cambios estructurales en el comportamiento de las transacciones.

En la Figura 1 (Tendencia), se observó un crecimiento casi lineal en el volumen de transacciones durante el periodo inicial. No obstante, hacia finales de marzo e inicios de abril de 2025 se evidenció un punto de inflexión, coincidente con una decisión estratégica tomada por Bancolombia (referencia). A partir de ese momento, la tendencia presentó un comportamiento más estable, lo cual sugiere una fase de consolidación o estabilización operativa.

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto., Imagen

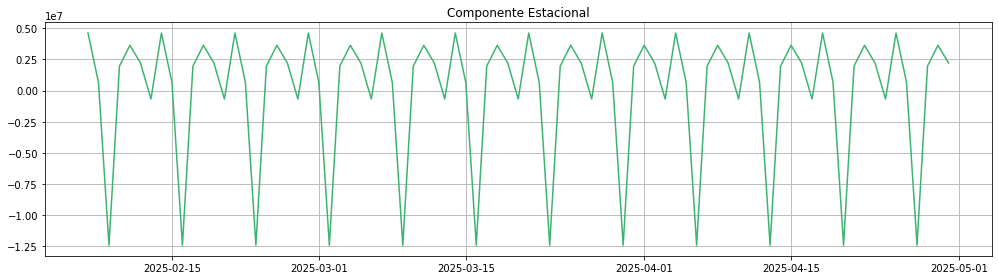
**Figura 1: Tendencia**

A partir del mes de abril, se identificó un aumento significativo en las transacciones, que se mantuvo con relativa constancia. Sin embargo, se observaron caídas puntuales en días específicos, especialmente a mediados de abril, que coinciden con fechas especiales como semana santa o puentes festivos. Este comportamiento refleja dinámicas conocidas por el negocio y, por tanto, no fue objeto de prueba en este estudio. Igualmente, se destacó un aumento en los días de quincena, lo que refuerza el impacto de los ciclos salariales en el volumen transaccional.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto., Imagen**Figura 2:** **Promedio de transacciones por día del mes**

En la Figura 3 (Estacionalidad), se identificó un patrón claramente semanal. Para cada mes analizado, se evidenciaron cuatro estaciones regulares, lo que resalta la necesidad de considerar los días de la semana como una variable clave en futuros modelos. Esta estacionalidad semanal es de gran relevancia para la caracterización de transacciones anómalas y la optimización del monitoreo operativo.



**Figura 3: Estacionalidad**

Finalmente, la Figura 4 (Función de Autocorrelación – ACF y PACF) evidenció una correlación significativa cada siete unidades de tiempo, lo que confirma la presencia de un comportamiento periódico de tipo semanal. Esta información respaldó la decisión de incorporar una variable categórica asociada al día de la semana en el diseño del modelo, con el fin de capturar mejor la dinámica repetitiva del sistema transaccional.

Imagen que contiene Histograma

El contenido generado por IA puede ser incorrecto., Imagen

Escala de tiempo

El contenido generado por IA puede ser incorrecto., Imagen

**Figura 4: Autocorrelación (ACF y PACF)**

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto., Imagen

**Figura 5: Promedio de transacciones por día de semana**

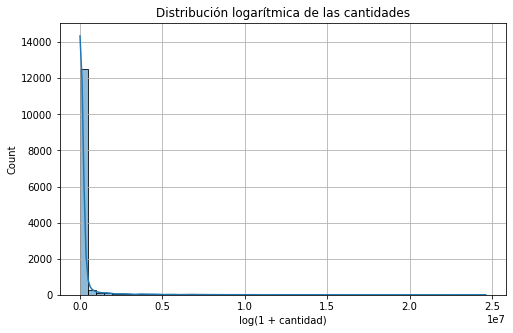
En el análisis transaccional realizado, se identificaron patrones diferenciados en el comportamiento de las transacciones a lo largo de la semana. En términos generales, los domingos se caracterizan por un volumen significativamente bajo de transacciones, lo cual es esperado debido a la menor actividad bancaria en ese día.

Por otro lado, los días laborales (lunes a viernes) muestran un comportamiento relativamente homogéneo en cuanto a volumen de transacciones. Sin embargo, es importante destacar que existen diferencias notables en algunos días específicos, principalmente cuando coinciden con eventos recurrentes que alteran el flujo normal de operaciones.

Las principales variaciones se observan cuando un día laboral corresponde a una quincena, lo que genera un aumento en el volumen de transacciones debido a los pagos de salarios. De manera similar, los días festivos también impactan el comportamiento, ya que las transacciones pueden concentrarse en ciertos momentos del día o incluso disminuir en otras franjas horarias.

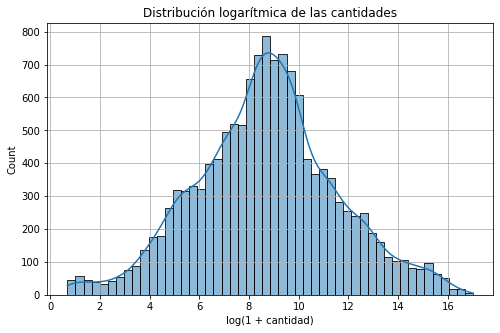
Además, se identificaron días atípicos que requieren un análisis diferente debido a situaciones excepcionales, tales como incidentes en la infraestructura del banco o situaciones externas que afectan la disponibilidad de los canales transaccionales.

**Interpretación estadística:**

****

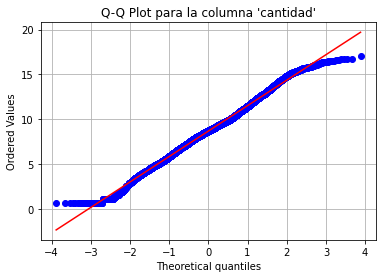
**Figura 6: Distribución volumen Transaccional**

Los datos originales presentan un sesgo positivo, caracterizado por una cola larga hacia la derecha con valores atípicamente altos. Este comportamiento probablemente se debe a eventos puntuales como días de afectación en los que los clientes realizan múltiples reintentos, jornadas de alto volumen transaccional como las quincenas, o incluso campañas de mercadeo que generan aumentos temporales en ciertas transacciones.



**Figura 7: Distribución logarítmica cantidades**

Como se observa en la figura 7 para corregir esta asimetría, se aplicó una transformación logarítmica, que reduce el impacto de los valores extremos y aproxima la distribución a una forma más simétrica. Tras esta transformación, se aplicaron pruebas de normalidad que confirmaron que los datos transformados siguen una distribución aproximadamente normal. Este enfoque es común en el análisis de datos financieros o transaccionales, donde unos pocos valores muy altos pueden distorsionar significativamente la distribución general.



**Figura 8: Q-Q Plot para la cantidad transaccional.**

Adicionalmente, contar con una distribución aproximadamente normal nos permite aplicar herramientas de estadística inferencial con mayor solidez. Por ejemplo, podemos calcular estimaciones más precisas de parámetros clave como la media poblacional (por ejemplo, el promedio diario de transacciones), lo que a su vez facilita la construcción de intervalos de confianza. Estos intervalos pueden convertirse en una base útil para generar alertas automáticas cuando los valores diarios se desvíen significativamente del comportamiento esperado.



**Figura 9: Resultados test de hipótesis de normalidad**

Las principales variaciones se observan cuando un día laboral corresponde a una quincena, lo que genera un aumento en el volumen de transacciones debido a los pagos de salarios. De manera similar, los días festivos también impactan el comportamiento, ya que las transacciones pueden concentrarse en ciertos momentos del día o incluso disminuir en otras franjas horarias. Además, se identificaron días atípicos que requieren un análisis diferente debido a situaciones excepcionales, tales como incidentes en la infraestructura del banco o situaciones externas que afectan la disponibilidad de los canales transaccionales.

# Preparación de los Datos

Como parte del proceso de preparación y comprensión de los datos, se realizó una extracción del histórico transaccional excluyendo días con comportamientos atípicos. Se definieron como días atípicos aquellos en los que se presentaron alertas significativas en la operación de los canales, así como fechas en las que se registraron incidentes tanto del lado del cliente como de la infraestructura del banco. La exclusión de estos registros respondió a la necesidad de garantizar un análisis más preciso del comportamiento transaccional habitual, reduciendo el ruido que podrían generar dichos eventos anómalos.

**Identificación de Transacciones de Bajo Volumen**

Para este análisis, se utilizó la técnica del rango intercuartílico (IQR) como mecanismo estadístico para identificar y filtrar las transacciones que se ubican en el extremo inferior de la distribución de frecuencias. Es decir, aquellas cuyo volumen operativo se encuentra significativamente por debajo del rango típico observado, lo que las clasifica como transacciones de baja volumetría. Este tipo de comportamiento puede estar asociado a actividades residuales, operaciones técnicas, eventos de baja concurrencia, o incluso etapas de transición en la dinámica del canal transaccional, como despliegues tecnológicos, horarios no convencionales o contingencias temporales.

Este filtrado permite aislar con precisión aquellas transacciones que, si bien no representan una carga operativa en términos de frecuencia, poseen un carácter estratégico o crítico dentro del ecosistema financiero, ya sea por el tipo de cliente involucrado, el valor de la operación o su impacto potencial en la continuidad del servicio. Como resultado, se obtiene un conjunto específico de transacciones dominadas por comportamientos de baja densidad de datos, que comúnmente quedan fuera del radar de los sistemas tradicionales de monitoreo, y que por lo tanto, representan el objetivo ideal para la aplicación de modelos de Machine Learning orientados a la detección temprana de anomalías.

Este enfoque permite enfocar los esfuerzos analíticos sobre un universo de operaciones que, aunque pequeñas en volumen, poseen una alta sensibilidad operativa y reputacional para el banco, facilitando la anticipación a fallas críticas y contribuyendo a la mejora continua de la disponibilidad y confiabilidad del servicio.

**Nuevas Características temporales**

Con el objetivo de obtener variables atributos relevantes temporales se divide del dataframe **df\_bajo\_volumen**, el cual incluye la fecha (dia), el tipo de transacción (transactioncodeop) y la cantidad. Para enriquecer el análisis, se generaron las siguientes variables adicionales:

**dia\_semana\_num**: Día de la semana como número (0=lunes, ..., 6=domingo), para identificar patrones semanales.

**es\_fin\_de\_semana**: Variable dicotómica (0/1) que indica si el día corresponde a sábado o domingo.

**dia\_mes**: Día del mes, extraído para análisis de distribución mensual.

**es\_quincena**: Variable dicotómica que identifica si el día corresponde a fechas cercanas a los ciclos típicos de pago en Colombia (1, 2, 14, 15, 16, 29, 30, 31).

**es\_festivo**: Variable dicotómica que indica si el día fue feriado, con base en el calendario oficial de festivos en Colombia.

Estas variables permiten segmentar el comportamiento y realizar análisis estadísticos detallados sobre el impacto que tienen estos factores temporales en las transacciones de bajo volumen.

**Análisis Estadístico y Visual**

**a) Día de la semana**

Se realizó un análisis de la media y distribución de transacciones según el día de la semana. Los gráficos de barras y boxplots mostraron que, aunque los valores promedio se mantienen relativamente constantes entre semana, existen valores atípicos más altos en días como miércoles o viernes, lo que sugiere que, incluso en bajo volumen, ciertos días presentan una mayor dispersión.

**b) Quincena vs No Quincena**

Se analizó la diferencia en el comportamiento durante fechas de quincena y fin de mes, momentos en los que se concentran los pagos salariales. Se observó un leve aumento en la media de transacciones en días identificados como quincena, lo cual refuerza la influencia de los ciclos de pago en el comportamiento financiero del usuario, incluso en canales o productos de baja volumetría.

**c) Festivos**

Se evaluó si los días festivos tenían un comportamiento distinto en cuanto al volumen de transacciones. En general, se evidenció una reducción en la actividad durante estos días, lo cual concuerda con la expectativa de menor uso de servicios financieros en fechas no laborales.

**d) Fines de semana**

Finalmente, el análisis evidenció una baja actividad relativa durante fines de semana, con menor cantidad promedio de transacciones. Esto indica que la mayoría de las transacciones de bajo volumen ocurren entre semana, probablemente vinculadas a operaciones corporativas o técnicas programadas en horario hábil.

**Conclusiones del Análisis**

El análisis de transacciones de bajo volumen permitió identificar que, aunque el volumen individual sea reducido, el comportamiento sigue patrones claros según el calendario laboral y financiero en Colombia. La concentración de transacciones alrededor de fechas clave como quincena, fines de mes y días hábiles evidencia que incluso las transacciones pequeñas responden a estructuras macroeconómicas y organizacionales.

Esta comprensión resulta valiosa para el diseño de modelos predictivos, sistemas de alertas y estrategias de gestión operativa más sensibles a los comportamientos sutiles, pero consistentes, del sistema transaccional.

# Parte 2

# Modelado (Propuesto)

Para el modelado se propone integrar un modelo de aprendizaje supervisado y uno no supervisado, tales como Decision Tree algorithm y K-Means Clustering, que pueden aportar un enfoque híbrido interesante para mejorar la detección de anomalías con datos reducidos. También se propone implementar un modelo de SVM (Support Vector Machine) el cual puede aportar sencillez, flexibilidad y bajo riesgo de sobre ajuste para datos reducidos.

**¿Por qué se selecciona cada modelo o técnica propuesta?**

Se elige Decision Tree por su capacidad para manejar conjuntos de datos pequeños, su interpretabilidad y su habilidad para capturar reglas de decisión claras. K-Means Clustering se incluye para explorar estructuras ocultas en los datos sin etiquetas, lo que ayuda a descubrir patrones anómalos o comportamientos atípicos. Por su parte, SVM se propone debido a su eficacia con datos de alta dimensión y escaso volumen, así como por su habilidad para mantener un rendimiento equilibrado al separar clases de forma óptima.

**¿Cuáles son las ventajas y limitaciones que se anticipan para cada técnica?**

Decision Tree: Ventajas incluyen su interpretabilidad y rapidez; sus limitaciones pueden ser el sobreajuste si no se ingresan eligen las variables realmente importantes para el modelo y la sensibilidad a pequeñas variaciones en los datos, .

K-Means: Tiene la ventaja de ser simple y eficiente para datos estructurados, pero es sensible a la selección del número de clusters y a los valores atípicos.

SVM: Presenta buen rendimiento en espacios de alta dimensión y un bajo riesgo de sobreajuste, aunque su entrenamiento puede ser costoso computacionalmente y menos interpretable que un árbol de decisión.

**¿Qué métricas de desempeño se usarán para evaluar inicialmente los modelos?**

Para los modelos supervisados como Decision Tree y SVM, se emplearán métricas como precisión, recall, F1-score y matriz de confusión. Para K-Means, se utilizarán indicadores como el coeficiente de silueta, la distancia intra-cluster y, cuando sea posible, la validación cruzada con etiquetas generadas manualmente o heurísticas externas

# Evaluación (Propuesta)

Para evaluar rigurosamente la solución desarrollada, es fundamental considerar su alineación con el objetivo principal: generar alertas automáticas ante posibles eventos de indisponibilidad en el servicio, especialmente en transacciones críticas de bajo volumen. La evaluación del modelo no solo se centrará en métricas estadísticas, sino también en su impacto operativo y utilidad en el contexto del negocio.

Uno de los principales riesgos identificados es el sobrealertamiento, es decir, la generación excesiva de alertas falsas que pueden saturar a los equipos de monitoreo y disminuir la confianza en el sistema. Por tanto, un criterio clave de éxito será lograr un equilibrio entre sensibilidad (detectar la mayor cantidad posible de eventos reales) y precisión (minimizar las falsas alarmas).

Para ello, se utilizarán métricas como:

Tasa de verdaderos positivos (recall): para asegurar que los eventos relevantes sean detectados.

Tasa de falsos positivos (false positive rate): para controlar el sobrealertamiento.

Precisión (precision): para evaluar qué proporción de las alertas generadas realmente correspondió a eventos de incidencia válidos.

F1 Score: como métrica balanceada entre precisión y recall, útil especialmente en contextos de datos desbalanceados.

Además, se implementarán técnicas de validación cruzada con diferentes subconjuntos de los datos y periodos de tiempo, para asegurar que el modelo sea robusto y tenga la capacidad de generalizar adecuadamente a nuevos datos. También se evaluará su desempeño en condiciones reales mediante pruebas en entornos controlados antes de su despliegue definitivo.

En términos de negocio, el modelo será considerado exitoso si logra detectar de forma oportuna los eventos relevantes con una tasa de alertas falsas aceptable, contribuyendo a una respuesta más rápida ante fallos y a una mejor disponibilidad del servicio.

# Despliegue (Propuesto)

El despliegue del modelo va dirigido a los analistas del equipo de monitoreo COES Bancolombia integrando el servidor de aplicaciones oficial del banco, el Datalake y Power Bi como herramienta de visualización, el servidor se conectará al datalake para extraer los datos en tiempo real y enviar los datos al modelo como aplicación ,el resultado es una tabla del mismo datalake que posteriormente será conectado por medio de webhook aun canal de Microsoft Teams, y será el medio por el cual se visualizará el resultado del modelo con los datos de producción.